Vers un environnement de production et de validation de ressources lexicales sémantiques

Mikaël Morardo Éric Villemonte de La Clergerie
Alpage, INRIA / Université Paris 7,
Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
{mickael.morardo,eric.de_la_clergerie}@inria.fr

RÉSUMÉ

Nous présentons quelques composants d'un environnement pour la production, la visualisation et la validation de ressources lexicales (termes et relations). Au coeur se trouve un composant de construction de réseau lexical entre termes s'appuyant sur l'hypothèse distributionnelle de Harris appliquée aux dépendances syntaxiques produites par l'analyseur FRMG sur gros corpus. Un autre aspect important concerne l'utilisation d'une interface WEB pour la visualisation et la validation collaborative des ressources produites.

ABSTRACT

Towards an environment for the production and the validation of lexical semantic resources

We present some components of a processing chain for the creation, visualization, and validation of lexical resources (terms and relations). At the heart, we find a component for building lexical networks relying on Harris' distributional hypothesis applied on the syntactic dependencies produced by the French parser FRMG on large corpora. Another important aspect concerns the use of a WEB interface for the visualization and collaborative validation of the resulting resources.

MOTS-CLÉS : hypothèse distributionnelle, réseau lexical, extraction terminologique, dépendances syntaxiques, interface WEB collaborative.

KEYWORDS: distributional hypothesis, lexical network, term extraction, syntactic dependencies, collaborative WEB interface.

1 Introduction

Chaque domaine de spécialité possède un ensemble de concepts, souvent exprimés sous forme de termes, qui lui est propre. Pour améliorer les performances de nombreuses tâches (indexation par méta-données, recherche d'information, extraction d'information, aide à la rédaction, etc.), il est important de pouvoir identifier, organiser, et maintenir l'ensemble de ces concepts et de leurs relations, ce rapidement et à coût humain réduit. Nous pensons que cela passe par deux grands axes complémentaires que nous cherchons à concilier dans l'environnement ici présenté.

Le premier axe repose sur l'utilisation de techniques d'acquisition à partir de corpus textuels. Dans notre cas, nous exploitons des corpus analysés syntaxiquement avec l'analyseur FRMG. Les sorties produites permettent l'extraction de termes (multi-mots) et la constitution de réseaux lexicaux

entre ces termes, en s'appuyant, comme de nombreux prédécesseurs (Bourigault, 2002; Lin & Pantel, 2002), sur l'hypothèse distributionnelle de Haris (Harris, 1968). Cette hypothèse stipule que des mots sémantiquement proches apparaissent dans des contextes similaires et conduit donc à examiner les contextes (syntaxiques dans notre cas) dans lesquelles apparaissent les mots. Elle nous amené à formuler et tester une variation de l'algorithme de regroupement markovien (MCL) (van Dongen, 2000) permettant de rapprocher d'une part les termes, et d'autres part les contextes syntaxiques.

Le deuxième axe guidant le développement de notre environnement concerne la mise en place d'une interface de visualisation et de validation collaborative. Il paraît en effet encore irréaliste de produire automatiquement des ressources parfaites, également difficiles à évaluer, rendant nécessaire une expertise humaine. L'interface permet de facilement parcourir les réseaux de termes, naturellement matérialisés sous forme de graphes, en fournissant des éléments d'explication (dont des exemples) et des fonctionnalités d'édition pour des validations plus rapides.

La section 2 présente le matériel de départ pour nos expériences, utilisé en particulier pour de l'extraction de termes multi-mots (section 3) et la constitution de réseaux lexicaux traduisant la proximité sémantique (section 4). Des éléments d'analyse des résultats sont fournis en section 5 avant de présenter section 6 les grandes lignes de notre interface de visualisation et de validation, implantée dans le cadre de la plateforme Libellex qui ouvre la voie à des investigations plus poussées des ressources produites.

2 Corpus annotés syntaxiquement

Nous avons mené plusieurs expériences d'acquisition sur divers corpus, spécialisés ou non. Ainsi, `All` (740 Millions de mots) est un corpus composite et généraliste [1] qui regroupe, entre autres, Wikipedia (180Mmots), Wikisource (littéraire, 64Mmots), et des dépêches AFP (journalistique, 248Mmots). Nous exploitons actuellement un corpus juridique (470Mmots) découpé en plusieurs sous-domaines (fiscal, affaires, social, civil et comptable). Nous traitons aussi des corpus beaucoup plus petits, comme un de description botanique (flores) ou un de maintenance automobile. Ces corpus sont analysés syntaxiquement avec l'analyseur du français à large couverture FRMG (Villemonte de la Clergerie, 2010), intégré dans la chaîne de traitement linguistique ALPAGE.

Pour nos expériences, les résultats d'analyse sont fournis au format Passage (Vilnat *et al.*, 2010), moins riche que le format DepXML nativement fourni par FRMG, mais produit par au moins une dizaine d'analyseurs syntaxiques pour le français. Le format Passage s'appuie sur 6 types de *chunks* (GN, GP, NV, PV, GA [adjectifs], GR [adverbes]) et 14 types de dépendances entre formes et/ou *chunks*, que l'on peut facilement ramener systématiquement à des dépendances entre formes. À partir des résultats d'analyses, certains éléments récurrents sont collectés et comptés, en s'appuyant sur une adaptation de l'algorithme MapReduce (Dean & Ghemawat, 2004).

1. Le choix de ce corpus général s'explique par des raisons historiques et également pour s'intéresser à un cas plus difficile que celui des domaines spécialisés, qui restent néanmoins le champ d'action naturel pour ce genre de travaux.

3 Extraction terminologique

Reprenant des idées bien connues (Pazienza *et al.*, 2005), l'extraction de termes candidats multi-mots s'appuie en premier lieu sur la collecte des séquences de *chunks* correspondant (approximativement) au motif `(GN)(GR*GA|GP|PV)+` [2], les composants étant liés syntaxiquement (essentiellement par des relations MOD-N pour *modifieur de nom*). Les candidats sont classés en fonction de multiples critères comme leur fréquence, leur cohésion interne (calculée via une mesure d'information mutuelle point à point), leur autonomie, et leur diversité d'utilisation. L'autonomie exploite les dépendances pour vérifier qu'un terme peut jouer des rôles actifs (comme sujet et objet) et n'est pas systématiquement modifié (par des GP par exemple). La diversité part du constat que certaines séquences sont parfois des collocations apparaissant dans des phrases clichées, souvent largement identiques, ce que nous pénalisons. Les variants sont regroupés en fonction des lemmes sous-jacents, et certains candidats sont rejetés à cause d'une trop grande variabilité (provenant par exemple d'un lemme `_NUMBER` ou `_DATE` correspondant à une entité nommée).

Avec un filtrage minimum, nous obtenons de l'ordre de 100K termes sur les sous-corpus AFP et Wikipedia et 50K sur la partie *fiscale* du corpus juridique (145M mots). Les termes extraits viennent avec un ensemble de phrases illustratives et des informations statistiques. La figure 3 montre quelques exemples de termes issus de la partie `affaires`.

4 Regroupement sémantique

Le regroupement sémantique des mots simples et des termes extraits précédemment s'appuie sur la collecte des dépendances syntaxiques, retournées sous forme de triplets (gouverneur, relation, gouverné) comme par exemple le triplet (*se_asseoir*, *sur*, *chaise*).

Nous ignorons certaines relations Passage (comme la juxtaposition) et appliquons certaines modifications pour essayer de nous rapprocher du niveau sémantique. Ainsi, sans être exhaustif :

- les passifs sont réécrits pour que le sujet de surface devienne un objet, et le complément introduit par *par* un sujet profond ;
- la relation ATB-SO (attribut du sujet ou de l'objet) qui lie le verbe à l'attribut relie l'attribut directement au sujet ou à l'objet ;
- la relation SUJ-V pour Passage s'attache sur le premier verbe, qui peut être l'auxiliaire : nous retrouvons alors le vrai verbe ;
- les relations mettant en jeu un coordonnant sont redistribuées sur les éléments coordonnés ;
- les modifieurs prépositionnels (MOD-N ou CPL-V) donnent un triplet précisant la préposition ;
- les prépositions sont postfixées par = quand elles introduisent un groupe nominal sans article ;
- des triplets supplémentaires sont ajoutés pour certains cas d'ambiguïtés d'attachement prépositionnels. Ainsi, dans une expression comme *journal de bord du capitaine*, la désambiguisation propose le triplet (*bord*,*de*,*capitaine*) mais nous ajoutons le triplet indirect (*journal*,*de**,*capitaine*) avec la marque *. La motivation est que les erreurs d'attachement induites par ces triplets indirects vont se distribuer sur de nombreux gouverneurs mais que les bons attachements (sur la tête d'un terme multi-mots) vont s'accumuler. La marque permet de filtrer

2. le `GN` initial peut aussi être un `GP` sans sa préposition.

ces triplets, et éventuellement de leur appliquer une confiance plus faible.
– dans la lignée du point précédent, des triplets sont aussi ajoutés quand le gouverneur ou le gouverné est tête d'un terme multi-mots issu de la phase d'extraction. L'exemple précédent ajouterait donc un triplet entre *journal de bord* (si détecté comme un terme) et *capitaine*. Dans `affaire` pour le terme *procédure collective*, on trouve ainsi 6462 occurrences de (*ouverture*,*de*,*procédure collective*), ou encore 189 occurrences pour (*qualité*,*de*,*président du conseil*).

La table 1 montre quelques exemples fréquents de triplets de dépendances, issus de `All`, impliquant le mot *chaise*. On constate la fréquence forte de la coordination ainsi que la forte représentation de triplets traduisant des termes (*chaise longue*, *chaise de poste*, *chaise musicale*, *chaise à porteur*, ...). Les triplets que l'on considère comme représentatifs de *chaise*, comme *s'asseoir sur*, bien que présents, sont loin d'être les seuls à forte fréquence.

gouverneur	relation	gouverné	fréq.	gouverneur	relation	gouverné	fréq.
chaise_nc	et	table_nc	**235**	prendre_v	cod	chaise_nc	**87**
asseoir_v	sur	chaise_nc	**227**	chaise_nc	modifieur	électrique_adj	**82**
chaise_nc	modifieur	long_adj	**168**	chaise_nc	modifieur	vide_adj	**80**
chaise_nc	de=	poste_nc	**115**	chaise_nc	à=	porteur_nc	**80**
tomber_v	sur	chaise_nc	**103**	dossier_nc	de	chaise_nc	**78**
chaise_nc	modifieur	musical_adj	**102**	avoir_v	cod	chaise_nc	**71**
se_asseoir_v	sur	chaise_nc	**93**	table_nc	et	chaise_nc	**62**

TABLE 1 – Quelques exemples de triplets de dépendance impliquant *chaise*.

Un triplet comme (*se_asseoir*,*sur*,*chaise*) permet d'associer le contexte <*se_asseoir sur* •> à *chaise* et le contexte <• *sur chaise*> à *se_asseoir*. Chaque mot w_i se voit ainsi associé un vecteur $[c_1 : u_{1i}, \cdots]$ de contextes c_a (avec leurs décomptes u_{ai}) qui vont servir à calculer la similarité. Comme ces vecteurs peuvent avoir un très grand nombre de composants, nous éliminons les contextes non lexicalisés par un mot plein (faisant par exemple référence à un pronom, comme <• *cln sujet*>), les contextes associés à un unique mot, et ceux à trop forte fréquence. De plus, les arguments phrastiques (correspondant à des complétives ou des infinitives) comme <• *cod manger*> sont réécrits en <• *cod* `*sentence*`>, considérant que, en première approximation, les verbes ainsi enchâssés ne sont pas réellement pertinents pour le regroupement de mots. Finalement, les mots à très faible fréquence (< 10 par défaut) sont éliminés.

Les paires (*mot* w_i,*contexte* c_a) fournissent les arcs d'un graphe biparti sur lequel nous adaptons l'algorithme de *clustering* markovien (MCL) (van Dongen, 2000). Celui-ci considère que des mots sémantiquement proches doivent être liés par un ensemble dense de chemins courts dans un graphe connectant les mots. L'algorithme originel s'appuie sur une matrice de contingence *mot-mot* et applique un opérateur d'*inflation/normalisation* tendant à renforcer la part relative des chemins courts et à diminuer celle des chemins longs. Dans notre cas, nous souhaitons préserver la dualité entre mots et contextes avec notre graphe biparti liant des mots w_i à des contextes c_a par des arcs orientés de poids $w_i \bullet c_a$ de w_i vers c_a et de poids $c_a \bullet w_i$ de c_a vers w_i. L'objectif est alors de connecter les mots par des poids $w_i \bullet w_j$ dénotant leur similarité et, similairement, de connecter les contextes par des poids $c_a \bullet c_b$, comme illustré par la figure 1.

Les formules (1) traduisent ces idées, en forme développée à gauche et en forme matricielle compacte à droite avec $W = (w_i \bullet w_j)$ la matrice de similarité sur les mots et $C = (c_i \bullet c_j)$ celle sur les contextes.

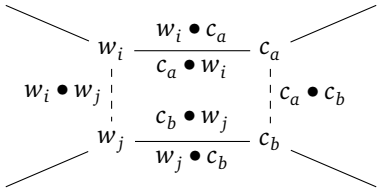

FIGURE 1 – Graphe termes – contextes pour l'algorithme de regroupement

$$\begin{cases} w_i \bullet w_j = \dfrac{1}{Z_i}\left(\displaystyle\sum_{a,b}(w_i \bullet c_a)(c_a \bullet c_b)(w_j \bullet c_b)\right)^{\alpha} & W = \Gamma_\alpha(F^t C F) \\ c_a \bullet c_b = \dfrac{1}{Z_a}\left(\displaystyle\sum_{i,j}(c_a \bullet w_i)(w_i \bullet w_j)(c_b \bullet w_j)\right)^{\alpha} & C = \Gamma_\alpha(G^t W G) \end{cases} \tag{1}$$

où les Z_i et Z_a dénotent les facteurs de normalisation donnés par

$$\begin{cases} Z_i = \displaystyle\sum_j \left(\sum_{ab}(w_i \bullet c_a)(c_a \bullet c_b)(w_j \bullet c_b)\right)^{\alpha} \\ Z_a = \displaystyle\sum_b \left(\sum_{i,j}(c_a \bullet w_i)(w_i \bullet w_j)(c_b \bullet w_j)\right)^{\alpha} \end{cases} \tag{2}$$

La forme développée stipule que la similarité $w_i \bullet w_j$ se calcule en sommant les contributions des divers chemins allant de w_i à w_j en passant par des contextes c_a et c_b, ce qui, récursivement, résume la contribution $c_a \bullet c_b$ des chemins allant de c_a à c_b. Le facteur d'inflation $\alpha \geq 1$, appliqué à ces contributions, augmente la part relative (après normalisation) des chemins forts (normalement les plus courts passant par des arcs de poids élevé). Par défaut, nous fixons $\alpha = 2$ mais ce taux est en fait modulé par le taux de connectivité des parties du discours. [3]

Sous la forme matricielle compacte, l'**opérateur d'inflation** Γ_α combine inflation et normalisation tandis que la matrice F (et sa transposée F^t) regroupe les poids $(w_i \bullet c_a)$ du contexte c_a pour le mot w_i (similaire pour G sur les contextes). Comme nous considérons qu'un contexte c_a utilisé avec de nombreux mots est moins intéressant qu'un contexte plus spécifique de seulement quelques mots (mais pas d'un seul mot !), nous pondérons les poids $w_i \bullet c_a$ selon la formule suivante (similairement pour $c_a \bullet w_i$ dans G) :

$$w_i \bullet c_a = \frac{ln(u_{ai}) * \eta_a}{\sum_b ln(u_{bi}) * \eta_b} \quad \text{avec} \quad \eta_a = ln\left(\frac{\#\text{mots distincts}}{\sqrt{|\{w_j | u_{aj} > 0\}|}}\right) \tag{3}$$

Les deux formules de l'équation (1) sont mutuellement récursives et le calcul est approché à l'aide d'un algorithme itératif de point fixe en démarrant avec des matrices initiales $W^{(0)}$ et $C^{(0)}$. Nous obtenons ainsi une version de base de l'algorithme.

3. Nos expériences ont en effet montré que certaines parties du discours (comme les noms communs) sont moins connectables que d'autres (comme les adverbes).

Cependant, nous étendons ce modèle de base, pour pouvoir explorer diverses intuitions linguistiques et utiliser des informations externes.[4] Ainsi, comme les contextes sont structurés et enchâssent en fait un mot en leur sein, nous considérons que les contextes $r.w_i$ et $r.w_j$ construits à partir de 2 mots similaires w_i et w_j avec une même relation r (par exemple `sujet`) ont des chances d'être également similaires. Réciproquement, si $r.w_i$ et $r.w_j$ sont similaires, alors w_i et w_j ont des chances d'être similaires. Ceci peut se formaliser par l'équation (4), en utilisant un opérateur τ de transfert des contextes vers les mots, et, réciproquement, un opérateur ρ de transfert des mots vers les contextes, avec, pour chaque relation r, un coefficient de transfert β_r (à 0,2 par défaut) et une matrice de transfert $T_r = (\langle w_i \bullet c_a \rangle_r)_{ia}$ avec $\langle w_i \bullet c_a \rangle_r = 1$ si $c_a = r.w_i$ et 0 autrement.

$$\begin{cases} W = \Gamma_\alpha(F^t C F + \tau(C)) \text{ avec } \tau(C) = \sum_r \beta_r T_r^t C T_r \\ C = \Gamma_\alpha(G^t W G + \rho(W)) \text{ avec } \rho(W) = \sum_r \beta_r T_r W T_r^t \end{cases} \quad (4)$$

Par ailleurs, l'équation (1) n'assure pas l'auto-similarité d'un mot avec lui-même, même partant de matrices initiales avec $w_i \bullet w_i = 1$. Cependant, il est possible d'assurer cette auto-similarité en réinjectant à chaque itération un bonus multiplicatif dans le calcul de $w_i \bullet w_i$. De manière générale, ceci s'exprime par la formule (5), avec utilisation de matrices $L = (l_{ij})$ et $K = (k_{ab})$ de bonus de similarité, pour lesquelles $l_{ij} = (i == j)$ et $k_{ab} = (a == b)$. Dans cette formule, I désigne la matrice identité et $\circ$ le produit point à point de Hadamard avec $(A \circ B)_{ij} = A_{ij}.B_{ij}$. Après application des bonus, les matrices sont renormalisées, mais sans inflation (Γ_1).

$$\begin{cases} W = \Gamma_1((I + L) \circ \Gamma_\alpha(F^t C F + \tau(C))) \\ C = \Gamma_1((I + K) \circ \Gamma_\alpha(G^t W G + \rho(W))) \end{cases} \quad (5)$$

Cette formulation générique permet d'étendre la notion de bonus à d'autres cas que l'auto-similarité. Ainsi, nous attribuons de tels bonus, plus ou moins importants, dans les cas suivants :

- en présence de coordinations entre w_i et w_j, dépendant du nombre d'occurrences ;
- en cas de typographie proche, par exemple pour des variations portant sur des diacritiques ;
- en cas d'équivalence modulo certains préfixes et suffixes, pouvant traduire une relation de morphologie dérivationnelle ;
- en cas d'utilisation d'une ressource germe fournissant des indications de similarité, par exemple une ressource de type wordnet.

Ces matrices sont également utilisées pour insérer des malus, par exemple entre mots de catégories syntaxiques incompatibles. Nous pénalisons aussi les rapprochements entre mots trop disparates en terme de fréquence ou de diversité de contextes. En effet, nous considérons qu'un mot w_1 très fréquent et productif ne doit pas être rapproché directement d'un mot w_2 rare et spécifique, car il existe probablement un mot intermédiaire w_3 en terme de fréquence, proche de w_1 et w_2. L'espoir est de voir ainsi émerger une hiérarchisation entre les mots.

Pour une paire de mots (w_i, w_j), un des intérêts de la formule (1), toujours présent dans ses variantes, est de pouvoir estimer l'importance $\langle w_i \bullet w_j, c_a \rangle$ d'un contexte c_a intervenant dans le

4. ce qui est une des raisons motivant le développement d'un nouvel outil de regroupement de mots.

rapprochement de w_i et w_j avec

$$\left\langle w_i \bullet w_j, c_a \right\rangle = \sum_b (w_i \bullet c_a)(c_a \bullet c_b)(w_j \bullet c_b) \tag{6}$$

Nous ne conservons alors que les contextes les plus significatifs expliquant (partiellement) la similarité $w_i \bullet w_j$. Il devient aussi possible d'ordonner ces contextes et de rechercher une corrélation entre le classement $\mathscr{C}_{ij}$ ainsi obtenu et le classement originel $\mathscr{D}_i$ des contextes pour w_i. Si $\mathscr{C}_{ij}$ est largement inclus dans le bon ordre dans $\mathscr{D}_i$ avec des poids assez similaires, alors on peut supposer que le rapprochement de w_j avec w_i ne remet pas en cause l'importance relative des contextes pour w_i. Nous mesurons cette similarité entre classements avec la formule (7), où rank_{ai} dénotant le rang de c_a dans $\mathscr{D}_i$ et rank_{aij} son rang dans $\mathscr{C}_{ij}$.

$$\Delta_{ij} = \sum_{c_a, w_i \bullet c_a > 0} \left(1 + \left|1 - \frac{\text{rank}_{aij}}{\text{rank}_{ai}}\right|\right)^{-1} \left(1 + \left|1 - \frac{\left\langle w_i \bullet w_j, c_a \right\rangle}{w_i \bullet c_a}\right|\right)^{-1} \tag{7}$$

Pour des raisons d'efficacité, à chaque itération, nous ne gardons, pour un mot w_i, que les mots w_j dont le niveau de similarité $w_i \bullet w_j$, modulé avec Δ_{ij}, dépasse un certain seuil. Par défaut, nous requérons que $\Delta_{ij} w_i \bullet w_j > 0.01$.

Toujours pour des raisons d'efficacité, nous ne conservons dans $\mathscr{D}_i$ pour w_i que les contextes c_a dont l'importance relative $w_i \bullet c_a$ dépasse un certain seuil ($0,2 \max_a(w_i \bullet c_a)$ par défaut). Dans les cas où les parts relatives sont très uniformes, nous ne conservons que les 50 premiers. Le mauvais aspect de ce choix est de focaliser la recherche de similarité entre mots sur ces au plus 50 contextes (qui diffèrent pour chaque mot), ce qui induit certains effets pernicieux dans le cas de mots très productifs en tant que constructeurs de termes. On a alors des contextes forts découlant des termes en question (comme <• *de terre*> pour *pomme*) qui ne sont cependant pas vraiment significatifs pour le mot en question (comme *pomme*). Pour remédier à cette situation, avant d'éliminer des contextes pour w_i, nous construisons un vecteur ri_i indiquant la présence (1) ou l'absence (0) dans les contextes de w_i de 1000 contextes initialement choisis aléatoirement, selon les principes du *Random Indexing* (Kanerva *et al.*, 2000). Ces vecteurs offrent une vue plus large sur les contextes des mots. En cas de rapprochement de w_i et w_j, un bonus/malus $b_{ij} = cos(r_i, r_j) - 1$ est ajouté dans la matrice L.

5 Analyse des résultats

En sortie de l'algorithme, nous obtenons une liste ordonnée de paires non symétriques de mots, avec, pour chaque paire (w_i, w_j), des informations sur la force du rapprochement et une liste de contextes significatifs c_a ordonnés selon $\left\langle w_i \bullet w_j, c_a \right\rangle$. Pour le corpus *cpl* [5], nous obtenons ainsi 19 960 mots w_i et 51 980 paires (w_i, w_j), ce qui représente déjà un nombre important de connexions. En symétrisant ce graphe orienté, nous obtenons un graphe non orienté de 20 922 nœuds et 47 065 arêtes. Pour le corpus `affaire`, nous obtenons 10 223 noeuds et 13 584 arcs. [6]

5. pour *Corpus Passage Long* (495Mmots), correspondant à `all` moins la sous-partie AFP.
6. le nombre d'arêtes est a priori quadratique en le nombre de noeuds, avec ici des ratios a^2/n de 0,010 et 0,011.

La table 2 donne des indications sur la connectivité en fonction de la catégorie syntaxique pour divers corpus. On constate que la majorité des mots connectés sont en fait des entités nommées, mais que cela ne représente qu'une infime partie de celles-ci (dû au fait que l'immense majorité des entités ont moins de 10 occurrences). Dans une moindre mesure, la situation est similaire pour les noms communs. Par contre, les adverbes se connectent très facilement. En conséquence, l'algorithme d'acquisition a été légèrement modifié pour tenter de corriger (mais pas totalement) ces distorsions en modulant le facteur d'inflation en fonction de la catégorie syntaxique, avec par exemple une inflation de $\alpha + 4$ pour les adverbes et de $\alpha - 1$ pour les noms communs. On note que les termes sont légèrement moins connectables que les noms communs, ce qui pourrait s'expliquer par le fait qu'un terme a vocation à éliminer les formulations synonymes.

corpus	**partie du discours**	**#mots**	**#conservés**	**%c/m**
cpl	np	1 779 848	8 749	0,5
	nc	35 417	5 782	16,3
	v	11 224	2 480	22,1
	adj	10 198	3 108	30,5
	adv	2 693	776	28,8
fiscal	terme	50 479	4 981	9,9
	nc	18 760	1 976	10,5
	v	4 783	593	12,4
affaires	terme	65 138	5 142	7,9
	nc	23 506	2 095	8,9

TABLE 2 – Distribution de la connectivité sur quelques corpus et parties du discours.

En examinant la liste des paires, on constate que l'algorithme regroupe correctement certaines paires de mots correspondant à des erreurs typographiques comme *boîte* et *boite*. On retrouve également des regroupements correspondant à des variations orthographiques, comme *clé* et *clef*, *lis* et *lys*, ou *audio-visuel* et *audiovisuel*. Noté dans les premières versions de l'algorithme, ce phénomène a ensuite été renforcé par l'octroi d'un faible bonus en cas de proche distance d'édition (en particulier sur les diacritiques).

D'autre part, on observe une forte proportion de rapprochements s'appuyant partiellement sur le bonus de coordination (25,7% pour **cpl**). Dans nombre de cas, ce bonus ne vient cependant que conforter une décision motivée par des contextes, comme par exemple pour *chaise* et *divan*.

TULIP (Auber, 2003), un outil de visualisation de très grands graphes, permet d'avoir une vue plus globale du réseau. La figure 2 montre ainsi un fragment du graphe pour `all` reliant des parties du corps humain (principalement les os et muscles). TULIP permet déjà d'avoir une impression subjective mais utile de la qualité du réseau lexical. Il fournit aussi des informations précieuses sur la topologie globale et locale du réseau.

En particulier, on observe rapidement la présence de *buissons* de mots fortement interconnectés. En s'appuyant sur quelques considérations topologiques et des mesures standards dans les graphes[7], nous avons pu ainsi repérer ces buissons et en extraire environ 4000 classes (pour **cpl**), avec par exemple :

7. l'algorithme s'appuie sur la forte proportion de connections entre les fils N_i d'un noeud N (indiquant son appartenance à un buisson) et sur la diffusion de couleurs entre noeuds d'un même buisson pour en trouver les limites.

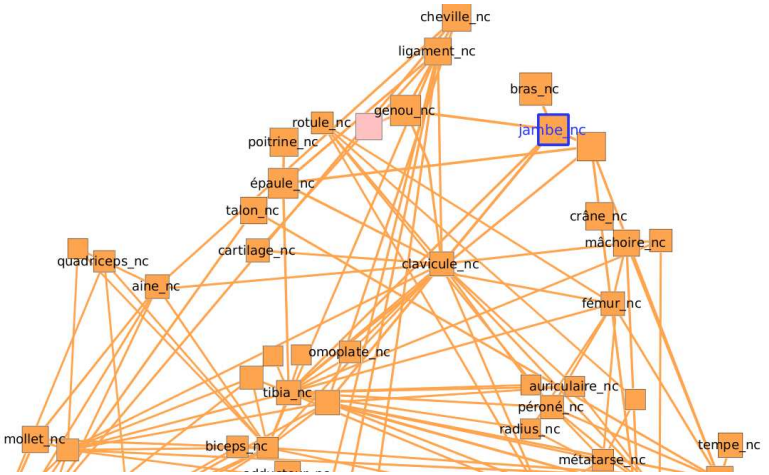

FIGURE 2 – Fragment de réseau, autour de *jambe*, visualisé sous le logiciel Tulip

<79> *sulky malinois fox-terrier setter cocker colley chiot fox labrador ratier griffon caniche teckel épagneul*

<80> *arrière-garde canonnier cavalerie carabinier tirailleur hussard panzer voltigeur blindé grenadier cuirassier avant-garde zouave lancier*

<83> *pneumonie paludisme diphtérie pneumopathie variole dysenterie malaria botulisme poliomyélite septicémie varicelle polio rougeole méningite*

Néanmoins, la vue avec TULIP reste subjective et ne dispense pas de tenter de mener des évaluations plus objectives. Nous en avons menées deux différentes, mais, pour des raisons de place, ne présentons ici que celle inspirée du test TOEFL [8] (Turney, 2002), en l'appliquant sur des ressources de type WordNet. Plus précisément, nous utilisons : [9]

– le WordNet du Français (**fwn**) (Jacquin *et al.*, 2007), qui ne couvre que les catégories nominales et verbales. Il comporte 22 121 synsets ;
– WOLF, un wordnet du français, aligné sur les synsets du wordnet anglais, automatiquement construit à partir de textes multilingues alignés et couvrant plus de catégories syntaxiques en ajoutant les adjectifs et les adverbes. Il comporte 32 351 synsets non vides regroupant 38 001 lexèmes distincts (Sagot & Fišer, 2008) ;

Un tel test TOEFL comprend une liste de questions, chaque question spécifiant un terme candidat et une liste de quatre termes réponses dont seulement un est proche sémantiquement du candidat. Le but du jeu est donc de retrouver la bonne réponse pour chaque question, avec une probabilité de réussite de 25% pour un programme répondant au hasard. N'ayant pas de test TOEFL sous la main pour le français, nous les engendrons à partir des wordnets, en choisissant le candidat et, aléatoirement, la bonne réponse (dans le même *synset*) ainsi que 3 autres termes leurres (dans d'autres *synsets*), comme illustré ci-dessous :

toutefois	*néanmoins*	complètement	progressivement	sensiblement
exploit	*prouesse*	offset	plie	bit

8. *Test of English as a Foreign Language*
9. Nous avons aussi mené des évaluations sur d'autres versions de WordNet qui nous omettons ici faute de place.

Néanmoins, on observe un faible recouvrement entre les vocabulaires des wordnets et du réseau. Nous filtrons donc les éléments du test, en demandant que les termes aient au moins 100 occurrences dans notre corpus, et ensuite pour une instance de réseau à évaluer, en demandant que les termes soient présents dans le réseau. Pour chaque réseau et wordnet, nous n'évaluons donc pas exactement sur le même jeu de test. En augmentant la taille des jeux, nous espérons néanmoins obtenir des mesures représentatives. Pour répondre aux tests, nous choisissons la réponse la plus proche de la question grâce à un chemin dans le réseau, et utilisons un tirage aléatoire en l'absence de chemins. Les résultats sont donnés par la table 3 pour plusieurs corpus sur les deux wordnets. Comme *baseline*, nous faisons aussi passer le test au wordnet WOLF, vu comme un réseau sémantique.

	fwn		**wolf**	
corpus	**%ok**	**#tests**	**%ok**	**#tests**
all	51,5	4 121	42,1	7 674
fiscal	46,1	104	37,0	493
affaires	35,1	248	43,2	1 055
social	39,4	274	37,7	1 345
wolf	64,5	1 076		

TABLE 3 – Évaluations de type Toefl.

On constate que les résultats sont en dessous de la référence fournie par le réseau wolf, ce qui n'est pas totalement surprenant, vu que WOLF a été certes automatiquement construit mais en s'appuyant totalement sur la structure du wordnet anglais. Les résultats sont néanmoins largement supérieurs à la *baseline* hasard de 25%. On note que les résultats sont meilleurs sur le French WordNet, mais avec une plus faible couverture en terme de vocabulaire (car FWN ne couvre que les catégories nominales et verbales). Pour les corpus spécialisés, les résultats sont, sans surprise, moins significatifs (moindre recouvrement du vocabulaire) et moins bons (en partie à cause d'une référence non spécialisée).

La table 4 fournit des résultats d'évaluation plus fins au niveau des catégories syntaxiques, la colonne **bons** (resp. **mauvais**) donnant le ratio de bonnes (resp. mauvaises) réponses obtenues en trouvant un chemin dans le réseau, tandis que **manquants** indique le ratio de cas où le choix de la réponse (bonne ou mauvaise) s'est faite par tirage au sort. On constate que, quand des chemins existent pour un `nc`, la réponse est presque toujours bonne (très forte précision) mais qu'on souffre de la faible connectivité des noms communs (faible rappel). À l'autre extrême, les adverbes sont facilement connectables, mais les réponses sont majoritairement fausses. Les verbes occupent une position intermédiaire.

wordnet	**pos**	**#tests**	**%bons**	**%faux**	**%manquants**	$\%b/(b+f)$
wolf	v	3 876	35,5	30,9	33,6	53,4
	nc	1 078	33,5	2,1	64,4	94,0
	adj	2 085	22,3	11,3	66,4	66,3
	adv	1 533	36,9	41,9	21,7	46,8

TABLE 4 – Tests Toefl par catégories syntaxiques (sur `cpl`).

6 Visualiser et valider avec Libellex

La situation à ce stade n'est pas totalement satisfaisante. Les vues fournies par un outil comme TULIP montrent la qualité de larges portions du réseau, tout en permettant de détecter, sans possibilité de corrections, un certain nombre d'erreurs. Les évaluations confirment que les réseaux ne sont pas parfaits mais démontrent aussi la difficulté de trouver des ressources de référence (il suffit de noter le manque d'accord entre French WordNet et Wolf). Le passage à des domaines spécialisés, sans référence existante et avec des réseaux de taille conséquente, pose aussi problème.

Dans le cadre d'une collaboration avec la société Lingua & Machina (L&M) pour étendre leur plateforme LIBELLEX, nous avons développé une interface Web pour parcourir, visualiser et valider des ressources lexicales, incluant termes et relations. Cette interface a pour objectif de répondre à plusieurs points essentiels :

– explorer facilement les résultats pour permettre aussi bien la mise en évidence de phénomènes dans la langue que d'erreurs dans les algorithmes.
– évaluer les résultats à la main tout en offrant une ergonomie facilitant la validation (les marquer comme acceptable ou non-acceptable avec une granularité variable).
– partager et échanger entre les utilisateurs sur des points ou des résultats ambiguës, difficiles à interpréter ou évaluer.
– conserver une trace des actions effectuées par un humain.

L'interface affiche une liste de termes (avec des options de tri et de recherche) avec les différentes informations associées au terme sélectionné (explications et phrases d'exemple). L'interface essaie d'être intuitive pour la navigation et les fonctions de validation (par exemple pour accepter ou rejeter un terme, changer le représentant vedette, enlever ou ajouter un variant, etc., pour ne parler que des termes) et doit faciliter le travail pour un validateur expérimenté ou débutant.

Le volume de données étant relativement grand (plus de 100K termes candidats pour la plus grosse ressource traitée dernièrement), travailler rapidement devient un impératif si l'on souhaite bénéficier dans un temps acceptable de données validées. L'interface offre la possibilité de communiquer entre utilisateurs pour partager leur expérience sur des points délicats (validité d'un terme par exemple) à l'aide d'un système de commentaires. En plus de cela, un utilisateur peut alerter sur des termes qu'il considère comme nécessitant une attention particulière.

Pour représenter les relations entre termes, la facilité de lecture qu'offrent les graphes a retenu notre attention, comme illustré par la figure 3. D'un coup d'oeil il est facile de repérer des anomalies comme des relations qui semblent dépourvues de sens ou des noeuds aux labels étranges. C'est pourquoi nous avons adopté un système de visualisation par graphe en s'appuyant sur la bibliothèque javascript `d3.js`. Pour ne pas surcharger l'utilisateur, seuls les voisins à deux générations du terme sélectionné sont visualisés, mais un click sur un noeud permet de recentrer la vue sur celui-ci. La taille des noeuds de seconde génération indique leur degré de connectivité. Il est également possible de rechercher les plus courts chemins entre deux termes.

Nous avons essayé de maintenir un graphe cohérent à l'affichage. Certaines parties d'un graphe peuvent contenir des sous-ensembles significatifs (des *clusters*) que souhaitons mettre en avant pour qu'ils soient identifiables au premier coup d'oeil. Ainsi, l'algorithme de placement des noeuds, qui s'appuie sur l'utilisation de forces attractives (intra-cluster) et répulsives (inter-cluster) permet de mieux séparer visuellement les divers sens qu'un terme possède (sous réserve

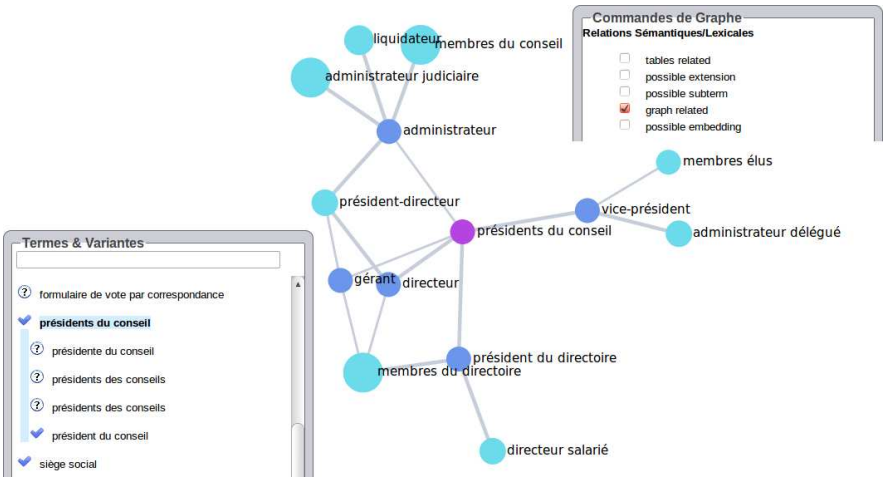

FIGURE 3 – Visualisation d'une partie du réseau sous Libellex (corpus `affaires`)

que l'algorithme d'extraction les ait trouvés). Les fonctionnalités d'édition permettent d'ajouter ou supprimer temporairement des relations (et prochainement de les valider).

Pour aider à prendre une décision lorsque l'on étudie le lien entre deux noeuds (ou plus), il est possible de consulter une matrice affichant, pour chaque terme sélectionné, les contextes les plus représentatifs ayant servi au rapprochement avec leur poids (indiqué par une jauge de couleur). Cette matrice fournit une vue synthétique et efficace pour vérifier la pertinence d'un lien entre deux noeuds. L'affichage de phrases d'exemples pour illustrer les contextes sert à lever toute ambiguïté. Il est à noter que cet outil a démontré son utilité dans de nombreux cas où l'intuition linguistique première d'un utilisateur était de considérer comme faux un lien donné.

Il est à noter que les matrices de contextes donnent de bons indices par rapport à la polysémie d'un terme. Ainsi, les mots *char*, *charrette* et *chariot* partageant les contextes <• *à bœuf*>, <• *modifieur atteler*>, <*promenade en* •> et d'autre part *char* et *tank* partageant les contextes <• *de combat*>, <*régiment de* •>, <• *modifieur lourd*>.

Nous pouvons donc vérifier, à différents degrés, si les données produites sont bonnes et, plus généralement, s'il n'y a pas d'erreurs dans les outils d'acquisition. En parcourant de manière locale les données, il devient plus simple de redonner du sens à des résultats statistiques mais il est aussi possible d'observer des phénomènes de langue plus difficilement identifiables avec d'autres formes de visualisation. Dans l'exemple de la figure 4, nous avons conclu que les contextes servant à rapprocher les termes sont, dans l'ensemble, liés au journalisme sportif. Il s'avère que la nature du corpus oriente les relations entre les termes. Il existe sûrement divers liens sémantiques entre *orteil* et *poignet*, mais, dans le corpus `all`, leur proximité est due à leur capacité à être le siège d'une blessure (*cassure*, *fracture*, etc.) pour des sportifs.

Sur un autre plan, un peu plus en marge, nous avons constaté l'intérêt que suscitent des graphes de mots dynamiques et interactifs au sein d'une population variée d'utilisateurs. C'est un phénomène intéressant à noter car démontrant un plaisir d'utilisation de la plateforme dans l'exploration des données : il y a un aspect ludique involontaire qui se dégage et qui mérite notre

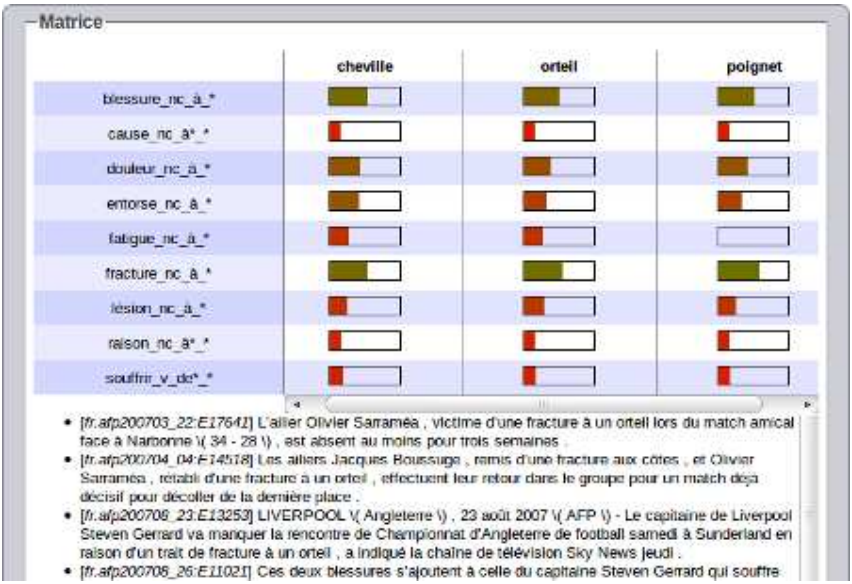

FIGURE 4 – Visualisation d'une matrice de contextes avec exemples

attention à l'avenir. En effet, éviter que l'utilisateur ne s'ennuie devant une tâche répétitive est un problème que nous considérons en permanence lors du développement. Nous avons donc une plateforme qui permet à la fois une exploration et une validation des données avec facilité et avec rapidité en plus d'un aspect collaboratif (à travers notamment le système de commentaires et de mise en avant d'un terme).

7 Conclusion

Nous avons présenté un environnement pour la production et la validation de réseaux lexicaux sur des termes simples ou multi-mots. Le coeur de cet environnement est un algorithme de regroupement de mots, flexible et capable de prendre en compte de nombreux paramètres, dont certains s'appuyant sur des intuitions linguistiques, comme par exemple la structure des contextes syntaxiques, les coordinations, des éléments de morphologie dérivationnelle et la distance d'édition.

Par ailleurs, bien que l'algorithme d'acquisition ait été conçu en prenant en compte la dualité des mots et des contextes, nous n'avons pas encore pleinement exploité ces derniers, même si l'interface de visualisation permet déjà de basculer sur une vue de contexte.

L'interface de visualisation et de validation collaborative, intégrée au sein de la plateforme Web LIBELLEX, nous semble un composant de plus en plus utile pour pouvoir mener des tâches d'évaluation plus précises (pour nous) et pour produire rapidement des ressources de qualité sur un domaine spécialisé (pour les utilisateurs). Comme l'expertise humaine est une chose précieuse, l'un des enjeux est de s'assurer que le travail d'un humain n'est jamais vain. Il doit d'abord être

conservé et non-altéré. Ensuite, nous devons l'utiliser pour des tâches d'évaluation et pour fournir du retour pour améliorer la chaîne de traitement (aux niveaux linguistique et acquisition). Enfin, les données validées doivent être exploitées dans les outils d'acquisition, par exemple sous forme de ressource germe dans la phase de regroupement de mots, ou pour entraîner des classifieurs pour mieux fixer les paramètres des algorithmes d'acquisition. On voit ainsi se mettre en place un cercle vertueux.

Il est à noter que nombre d'éléments de l'environnement sont remplaçables, que ce soit la chaîne linguistique, le système d'extraction de termes, ou celui de regroupement de mots. Le composant qui reste le plus indispensable est peut-être au final l'interface de visualisation et de validation.

Cette interface est accessible en ligne sur `http://alpage.inria.fr/Lbx` (avec l'identifiant `guest` et mot de passe `guest`) et nous sommes ouverts au chargement de nouvelles ressources, éventuellement produites à partir de notre chaîne de traitement à partir d'un corpus spécialisé.

Références

AUBER D. (2003). Tulip : A huge graph visualisation framework. In P. MUTZEL & M. JÜNGER, Eds., *Graph Drawing Softwares*, Mathematics and Visualization, p. 105–126. Springer-Verlag.

BOURIGAULT D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Proc. of TALN'02*, p. 75–84, Nancy, France.

DEAN J. & GHEMAWAT S. (2004). MapReduce : Simplified data processing on large clusters. In *OSDI'04 : Sixth Symposium on Operating System Design and Implementation*, San Francisco, CA.

HARRIS Z. (1968). *Mathematical Structures of Languages*. New-York : John Wiley & Sons.

JACQUIN C., DESMONTILS E. & MONCEAUX L. (2007). French eurowordnet lexical database improvements. In *In Proc. of CICLing'07*, number 4394 in LNCS, Mexico City, Mexico.

KANERVA P., KRISTOFERSON J. & HOLST A. (2000). Random indexing of text samples for latent semantic analysis. In L. GLEITMAN & A. E. JOSH, Eds., *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, p. p. 1036, Mahwah, New Jersey : Erlbaum.

LIN D. & PANTEL P. (2002). Concept discovery from text. In *Proceedings of Conference on Computational Linguistics (COLING-02)*, p. pp. 577–583, Taipei, Taiwan.

PAZIENZA M. T., PENNACCHIOTTI M. & ZANZOTTO F. M. (2005). Terminology extraction : an analysis of linguistic and statistical approaches. In S. (ED.), Ed., *Knowledge Mining*, volume 185 of *Studies in Fuzziness and Soft Computing*. Springer Verlag.

SAGOT B. & FIŠER D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. In *TALN 2008*, Avignon, France.

TURNEY P. D. (2002). Mining the web for synonyms : PMI-IR versus LSA on TOEFL. *CoRR*, **cs.LG/0212033**.

VAN DONGEN S. (2000). *Graph Clustering by Flow Simulation*. Phd thesis, University of Utrecht.

VILLEMONTE DE LA CLERGERIE E. (2010). Building factorized TAGs with meta-grammars. In *TAG+10 : The 10th International Conference on Tree Adjoining Grammars and Related Formalisms*, p. pp. 111–118, New Haven, CO.

VILNAT A., PAROUBEK P., VILLEMONTE DE LA CLERGERIE E., FRANCOPOULO G. & GUÉNOT M.-L. (2010). PASSAGE syntactic representation : a minimal common ground for evaluation. In *LREC*, La Vallete.